Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prognostic effect of inflammatory genes on stage I-III colorectal cancer – integrative analysis of TCGA data

Eun Kyung Choe
Seoul National University Hospital https://orcid.org/0000-0002-7222-1772
Sangwoo Lee
Cyber University of Korea
So Yeon Kim
Ajou University
Manu Shivakumar
University of Pennsylvania Perelman School of Medicine
Kyu Joo Park
Seoul National University College of Medicine
Young Jun Chai
Seoul National University Seoul Metropolitan Government Boramae Medical Center
Dokyoon Kim (dokyoon.kim@pennmedicine.upenn.edu)
https://orcid.org/0000-0002-4592-9564

Research article

Keywords: colorectal cancer, prognosis, Inflammatory status indicators

Posted Date: April 7th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-18854/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Cancers on February 11th, 2021. See the published version at https://doi.org/10.3390/cancers13040751.

Abstract

Background Inflammatory status indicators have been reported as a prognostic biomarker of colorectal cancer (CRC). However, since the inflammatory interactions with colon involve various modes of action, the biological mechanism to link inflammation and CRC prognosis is not fully elucidated. We comprehensively evaluated the predictive role of the expression and methylation level of inflammation-related genes for CRC prognosis and their pathophysiological associations.

Method An integrative analysis was conducted on 247 patients of stage I-III CRC from The Cancer Genome Atlas. Lasso-penalized Cox proportional hazards regression (Lasso-Cox) and statistical Cox proportional hazard regression (CPH) were used for analysis.

Result Models to predict overall survival were designed with respective combinations of clinical variables, including age, sex, stage, gene expression, and methylation. An integrative model combining expression, methylation, and clinical features had the highest performance (median C-index=0.756), compared to the model with clinical features alone (median C-index=0.726). By multivariate CPH with features from the best model, methylation levels of CEP250, RAB21 and TNPO3 were significantly associated with overall survival. They did not share any biological process in functional networks. The 5-year survival rate was 29.8% in a low methylation group of CEP250 and 79.1% in a high ($P < 0.001$).

Conclusion Our study result implicates the importance of integrating the expression and methylation information along with clinical information in prediction of survival. CEP250, RAB21 and TNPO3, in the prediction model might have a crucial role in CRC prognosis and further improve our understanding of potential mechanisms linking inflammatory reaction and CRC progression.

Introduction

The role of inflammation in the progression and prognosis of cancer is a growing interest in oncologic research[1]. Candidate inflammatory biomarkers such as neutrophils, lymphocytes, monocyte, platelets, lymphocyte to monocyte ratio (LMR), neutrophil to lymphocyte ratio (NLR), and platelet to lymphocyte ratio (PLR) have been suggested as a predictive factor for cancers prognosis [2-4]. In colorectum, molecular pathophysiology implicates inflammation would promote the progression of colorectal cancer [5]. The role of inflammation on colorectum is well-manifested in inflammatory bowel disease such as ulcerative colitis and Crohn's disease. In these patients, chronic inflammation is a major risk factor for the development of gastrointestinal malignancies, especially colorectal cancer (CRC) [6]. Inflammatory biomarkers such as NLR, LMR and PLR have been suggested as prognostic markers in colorectal cancer [7-10]. Chronic inflammation can trigger genetic mutations and epigenetic alternations that promote malignant cell transformation [11]. However, the inflammatory interactions with colon involve various modes of action such as immune cells, cytokines, and other immune mediators in virtually all sequence of CRC progression, including initiation, progression, and metastasis [12]. And in recent reports, it has been postulated that epigenetic changes have a critical role to establish efficient expression profiles

during inflammation and disease [13-15]. For this complexity of the association between inflammation and CRC, the underlying mechanism to link inflammation and CRC prognosis is not fully elucidated. To address this mechanism, integrative analysis of multi-omics data with a genome-wide view would be necessary [16].

The Cancer Genome Atlas (TCGA) has a collection of genomic and epigenomic data along with clinical data of patients for a large number of cancer types [17]. TCGA aims to develop catalog of major cancer genomics through integrated multi-dimensional analyses [18]. This purpose has been increasingly implemented to predict various oncological outcomes based on the multi-omics data [19-22].

In this study, we systematically evaluated the expression and DNA methylation level of the inflammation-related genes in CRC tissues using the TCGA-COREAD dataset. We comprehensively analyzed the whole set of inflammation-related genes from both RNA expression and methylation to evaluate the predictive role in overall survival of those genes and examine the significance level in association with overall survival. By using the Lasso-penalized Cox proportional hazards regression (Lasso-Cox), features were selected to design the prediction model for overall survival and used to evaluate whether the genetic features could improve the prediction performance additive to clinical prognostic factors such as age and TNM staging. To our knowledge, no previous studies have focused on the use of integrated omics biomarkers, particularly inflammation-related genes, as additive features to the basis of the current TNM staging in designing a prediction model for overall survival in colorectal cancer.

Methods

Data acquisition

RNA-seq data of The Cancer Genome Atlas-colorectal adenocarcinoma (TCGA-COREAD) was downloaded from the UCSC Xena browser (https://xenabrowser.net). A gene-level methylation data of TCGA-COREAD was retrieved from the Broad Institute Firebrowse (http://firebrowse.org). For the RNA-seq data, genes that have more than 50% samples with 0 expression levels were removed. For the methylation data, we used the methylation data by picking the probe with the minimum correlation with mRNA-seq data for each gene. Genes that have zero methylation levels in all samples were excluded. Among the 736 samples in the data, we excluded the samples based on the clinical information as below, samples collected from normal tissue; stage IV colorectal cancer; missing data on pathological stages, survival data, and age information; and follow-up duration less than 12 months. Finally, we chose the data for analysis using the overlapping sample IDs among RNA-seq, methylation and clinical data set. Thus, 247 patients were used for further analyses.

Ethics statement

On TCGA publication guidelines (http://cancergenome.nih.gov/publications/publicationguidelines), there are no restrictions on the publication of to use the dataset. The study was conducted in accordance with the Declaration of Helsinki.

Feature selection

Figure 1 shows an overview of the analysis framework. First, we selected the gene sets related to the inflammatory status from the National Center for Biotechnology Information's (NCBI) Gene database (www.ncbi.nlm.nih.gov/gene) [23], with the filtering term “(inflammatory) AND "Homo sapiens"[porgn:_txid9606]”. 2,576 genes were retrieved and used as inflammation-related genes for the analyses. Then, we preprocessed the gene sets and the clinical information (such as age (< 65years vs. >= 65years), sex (male vs. female), T stage (T1, T2 vs. T3, T4) and N stage (N0 vs. N1, N2) based on their significances of association with overall survival by performing univariate cox proportional hazard regression (CPH) analysis. Clinical, expression and methylation features that passed the suggestive significance level ($P < 0.01$) were used for further analyses (Clinical, Expression, Methylation). We also made an input matrix where Expression and Methylation were concatenated (Expression+Methylation) (Figure 1). Lastly, feature selection was performed for respective data sets (Clinical, Expression, Methylation, and Expression+Methylation) by using Lasso-penalized Cox proportional hazards regression (Lasso-Cox). The training set versus test set ratio was 8 : 2. We selected important features that show the best prediction performance by training the Lasso-Cox model in Clinical, Expression, Methylation, and Expression+Methylation dataset. To validate the feature selection model, we performed 50 iterations of 5-fold cross-validation in the training set.

Overall survival prediction based on omic features

Using the selected features, we trained prediction models for overall survival using Lasso-Cox and compared the performances of each model. Seven models were designed by respective combination of features such as Clinical features (C); Expression features (E); Methylation features (M); Expression + Methylation features (EM); Clinical features + Expression features (C-E); Clinical features + Methylation features (C-M); and Clinical features + Expression + Methylation features (C-EM). The performances among the models were compared by concordance index (C-index) which measures how concordant the observations and predictions by our model are for a pair of randomly chosen patient samples [24](Figure 1). Among the models, we used the model which achieved the highest C-index to perform multivariate CPH analysis to examine which features have a significant association with overall survival ($P < 0.05$). Then, Kaplan-Meier analysis was carried out using the genes which were significantly associated with overall survival. The optimal cut-off points for those genes to divide patients into two groups were determined by MaxStat packages in R (Maximally selected Rank Statistics). MaxStat performs a test of independence of response and one or more covariables using maximally selected rank statistics[25]. The log-rank test was used to compare the survival curves of the different gene-level groups. Kaplan Meier plot was visualized by “ggkm” R packages. Finally, Integrative Multi-species Prediction (IMP) [26] was used to visualize the gene-gene network and biological process network. IMP integrates multiple sources of evidence for functional interactions by integrating 3,741 genome-scale datasets for 7 organisms and predicts for 582 disease and 12,117 biological processes (http://imp.princeton.edu/)[27]. We queried IMP with the gene set from the best prediction model. All the statistical and computational analyses were done by R statistical software (version 3.5.3 R development Core Team; R Foundation for Statistical

Computing, Vienna, Austria). The putative association between clinical information and survival outcome were assessed by Chi-square test, Student's t-test and analysis of variance (ANOVA) for independent groups in Table 1.

Results

Patients demographics

Our study population comprised of 247 patients (131 males and 116 females) who had colectomy for stage I-III colorectal cancer. The mean patient age was 64.28 +/- 12.90 years, and the median follow-up was 26.83 months (ranging from 12.17 to 150.07 months). Patients characteristics are shown in Table 1.

Feature selection for overall survival with Lasso-Cox

We preprocessed the gene sets and the clinical information based on the significance of association with overall survival by performing univariate CPH analysis. Features that passed the suggestive significance level ($P < 0.01$) were 4 clinical variables (age, sex, T stage and N stage), 26 gene expression, and 14 methylation features, respectively. By concatenating the selected expression and methylation features, Expression+Methylation data had 40 features. After feature selection using Lasso-Cox, 4 variables were selected for clinical features (C), 16 expression features (E), 12 methylation features (M), and 14 omic features from the combined dataset (EM). The list of the features is shown in Table S1.

Training prediction models for overall survival with Lasso-Cox

Using various combinations from 4 feature sets (C, E, M, and EM), we designed seven models to predict overall survival (Figure 1). The performances of each model are shown in Table 2. Among 4 respective feature sets, a model with EM showed the highest performance (median C-index =0.727). In particular, among models incorporating the genetic features into clinical features, an integrative model with C-EM showed the best performance (median C-index =0.756) (Table 2). Adding the respective E or M to C did not improve the performance compared to the model with C alone, however, integrating the EM and C showed the great improvement in our prediction model.

Association analysis for overall survival with Multivariate cox proportional hazard regression

For the model with C-EM, which is the best prediction model, an association study was performed to examine the association between the selected features and overall survival. By multivariate CPH, CEP250 ($P = 0.035$), RAB21 ($P = 0.002$) and TNPO3 ($P = 0.011$) as methylation features showed independent associations with overall survival (Table 3). For the three methylation features, Kaplan-Meier plots were generated (Figure 2). The patients were divided by the optimal cut-offs for each feature based on MaxStat. (-0.923 for CEP250; 0.5 for TNPO3, and -0.367 for RAB21). Figure 2 shows the Kaplan-Meier plots with the results of log-rank test. The 5-year survival rate of TNPO3 was 81.1% for methylation below cut-off point and 76.5% for above; in RAB21, 54.4% for below and 76.5% for above; and in CEP250, 29.8% for below and 79.1% for above.

Gene-gene network and biological process network

We visualized the functional network of the expression and methylation features used in the prediction model C-EM using IMP. For a gene-gene networking construction, we queried all 14 genes used in the model and the results are shown in Figure 3(a). Except for RAB21, TERF21P and PPARGC1A, 11 genes among 14 genes had less than 3 strong relationships with other genes in the network. The biological process network for the 3 genes, which are significantly associated with overall survival, was also visualized (Figure 3(b)). We used the confidence level of 0.5 when visualizing networking. Notably, all genes had unique the biological process and there were no shared biological processes among three genes.

Discussion

Age, sex, T stage, and N stage that are used in this study as clinical features, are the most commonly used clinical factors to anticipate the prognosis of colorectal cancer survival [28]. Comparing the prediction performance of these clinical factors, the integrative model that integrates clinical, expression and methylation features significantly improved the performance of prediction for overall survival. Adding expression feature alone, or methylation feature alone to the clinical features did not contribute to the improvement in performances. However, only the combined dataset between expression and methylation feature was able to improve the performance. This could be the finding coming from the fact that inflammatory interaction with CRC includes a wide range of biological processes in immune cells, cytokines, and other immune mediators in virtually all sequences of CRC progression, including initiation, progression, and metastasis[12]. This result is also consistent with our results using IMP. In the gene-gene network (Figure 3(a)), the selected genes in the prediction model do not have high-dimensional networks among each other, except for RAB21 and PPARGC1A, all genes have less or no connections to other genes. This could imply that those genes are uniquely contributing to the inflammation-cancer progression mechanism. Additionally, in the biological process network (Figure 3(b)), the genes associated significantly with overall survival (RAB21, CEP250 and TNPO3), did not share any biological processes. This seems to be in accordance with the finding that they were statistically independent factors for association with overall survival.

In the best integrative prediction model, 7 expression and 7 methylation features were selected. We did a literature review on the function of those gene features and the relation with inflammation. CEP250 and TERF2IP were involved in the cell cycle. CEP250 encodes centrosomal protein contributing centrosome–centrosome cohesion during interphase of the cell cycle [29]. It has been reported to be associated with the inflammation in joints and bones [30]. There is also a genome-wide association study that shows the association of CEP250 with inflammatory bowel disease [31]. TERF21P encodes a protein involved in telomere length regulation [29]. Several studies shows their aberrant activation leads to increased cancer cell proliferation and tumorigenesis [32, 33].

RAB21 and NINJ1 are related to cell adhesion and migration. RAB21 is Rab family of monomeric GTPases, and the encoded protein is related to the regulation of cell adhesion and migration [23]. It is involved in the inflammatory reaction by regulating lipopolysaccharide-induced pro-inflammatory responses [34]. It has been reported to be related to cancer invasion and cancer mobility [35, 36]. NINJ1 produces a protein that is homophilic cell adhesion molecules [37] and plays a role in the progression of multiple sclerosis [38]. NINJ1 is a target of p53 and has been known to repress WT p53 expression [39].

TNPO3, MAZ and PPARGC1A are related to gene regulation. TNPO3 encodes nuclear import receptor for serine/arginine-rich (SR) proteins [29]. This gene is reported to be associated with several inflammatory autoimmune diseases such as systemic lupus erythematosus [40] and rheumatoid arthritis [41]. There are no reports related to colon disease or cancer. MAZ affects the expression of MYC [42]. MYC is well known for the role in the progression of CRC [43]. MYC is frequently deregulated in inflammation and the expression is affected by DNA-methylation [11]. PPARGC1A encodes a transcriptional coactivator that regulates the genes involved in energy metabolism [29]. In an experiment, it was profoundly reduced in ulcerative colitis patients [44].

There are several genes related to cytokine reaction. TNFRSF18 and TNFSF12 are related to tumor necrosis factor (TNF), which is a well-known inflammatory biomarker as a tumor-promoting cytokine in colon cancer [12]. NLRP14 encodes NLRP (Nucleotide-binding oligomerization domain, Leucine-rich Repeat and Pyrin domain containing) protein family, which belongs in the inflammasome, that activates the pro-inflammatory enzymes [29]. NLRP inflammasome is suggested as a possible linkage mechanism between obesity-associated low-grade chronic inflammation and CRC development [45]. PTGES encodes a glutathione-dependent prostaglandin E synthase and its expression seems to be induced by proinflammatory cytokine interleukin 1 beta (IL1B) [29]. PTGES is upregulated in CRC and premalignant lesions such as colonic adenomatous polyps [46, 47]. DEFA5 encodes a family of antimicrobial and cytotoxic peptides thought to be involved in host defense [29]. It is reported as a candidate biomarker for Crohn's disease [48]. PRG4 encodes a large proteoglycan [29]. It plays an important anti-inflammatory role in osteoarthritis synoviocyte proliferation [49]. In addition, PRG4 has also been demonstrated to have anti-inflammatory properties [50] and it may suppress breast cancer cell invasion [51]. There are no reports related to colon disease or cancer.

For TMEM184A, there are not so many reports on its pathophysiology for inflammation and cancer progression. Few reports show that TMEM 184A functions as a heparin receptor and regulates the anti-inflammatory response of endothelial cells [52].

As seen in the literature review, the selected inflammatory genes have a variety of pathophysiological mechanisms that could be considered to contribute to inflammation – CRC linkage mechanism. These genes could be the crucial sets of genes that could cover the majority of the mechanism linking inflammation - CRC progression. Also, this could be postulated that the role of inflammatory genes for CRC progression is not related or connected to one functional pathway but is affecting through multiple global ways of mechanisms by respective inflammatory genes.

This study has several advantages. We comprehensively analyzed the whole genes related to inflammation to see the global network between inflammation and CRC progression and also include both expression, methylation data for those genes. By combining the expression and methylation data in model design, we were able to get the survival prediction model which improved the performance of conventional clinical prediction models. This performance would not be achievable if the model is designed using either gene expression or methylation data. Second, we performed both the prediction model based on computational analysis and association study based on statistical analysis. This will help to interpret the results more intuitively and support evidence for the results of the prediction model.

However, there are a couple of limitations in our analysis. First, as we used an open-source database, there were a lot of missing data. This prohibited us from using important clinical factors such as microsatellite instability, lymphatic invasion, and venous invasion for survival analysis. Second, we simply concatenated the clinical factors, RNA expression and methylation level for integration. Transformational integration could be applied in a larger set of samples. Third, since there are no open-source databases that include both RNA expression and methylation data, we could not replicate the results in another data set.

Conclusions

By analyzing the colorectal cancer patients from TCGA data, we were able to develop prediction models for overall survival using various features. Overall survival can be predicted with the best performance by model combining expression and methylation features to clinical features, which outperforms the model designed by clinical features alone. Functionally, the gene features in this model rarely share functional network implying that the inflammatory status contribute to the colorectal cancer progression, not in a simple biological process but a global gene set linkage of biological processes. By association analysis, we were able to identify the methylation level of CEP250, RAB21, and TNPO3 independently played a crucial role in colorectal cancer prognosis. The findings could elucidate the underlying mechanism between inflammatory status and colorectal cancer progression.

Abbreviations

CRC : colorectal

TCGA-COREAD : the Cancer Genome Atlas-colorectal cancer

Lasso-Cox : Lasso-penalized Cox proportional hazards regression

CPH : cox proportional hazard regression

IMP : Integrative Multi-species Prediction

LMR : lymphocyte to monocyte ratio

NLR : neutrophil to lymphocyte ratio

PLR : platelet to lymphocyte ratio

C : Clinical features

E : Gene expression features

M : Gene methylation features

EM : concatenated expression and methylation features

Declarations

Funding

“This work was supported by the National Library of Medicine (NLM) R01 NL012535.”

Availability of data and materials

All of the data used in this study can be acquired from Firebrowse (http://firebrowse.org)

Conflict of interest

The authors declare that they have no conflict of interest.

Author contribution

Study conception and design was done by EKC, SL, SYK, MS, KJP, YJC, DK;

Acquisition of data was done by EKC, YJC; Analysis and interpretation of data was done by EKC, SL; Drafting of manuscript was done by EKC, SL, DK

References

1. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. Nature. 2008; 454(7203):436-444.
2. Sun K, Chen S, Xu J, Li G, He Y. The prognostic significance of the prognostic nutritional index in cancer: a systematic review and meta-analysis. J Cancer Res Clin Oncol. 2014; 140(9):1537-1549.
3. Templeton AJ, Ace O, McNamara MG, Al-Mubarak M, Vera-Badillo FE, Hermanns T, et al. Prognostic role of platelet to lymphocyte ratio in solid tumors: a systematic review and meta-analysis. Cancer Epidemiol Biomarkers Prev. 2014; 23(7):1204-1212.
4. Gu L, Li H, Chen L, Ma X, Li X, Gao Y, et al. Prognostic role of lymphocyte to monocyte ratio for patients with cancer: evidence from a systematic review and meta-analysis. Oncotarget. 2016;

7(22):31926-31942.
5. Janakiram NB, Rao CV. The role of inflammation in colon cancer. Adv Exp Med Biol. 2014; 816:25-52.
6. Axelrad JE, Lichtiger S, Yajnik V. Inflammatory bowel disease and cancer: The role of inflammation, immunosuppression, and cancer treatment. World J Gastroenterol. 2016; 22(20):4794-4801.
7. Chan JC, Chan DL, Diakos CI, Engel A, Pavlakis N, Gill A, et al. The Lymphocyte-to-Monocyte Ratio is a Superior Predictor of Overall Survival in Comparison to Established Biomarkers of Resectable Colorectal Cancer. Ann Surg. 2017; 265(3):539-546.
8. Li MX, Liu XM, Zhang XF, Zhang JF, Wang WL, Zhu Y, et al. Prognostic role of neutrophil-to-lymphocyte ratio in colorectal cancer: a systematic review and meta-analysis. Int J Cancer. 2014; 134(10):2403-2413.
9. You J, Zhu GQ, Xie L, Liu WY, Shi L, Wang OC, et al. Preoperative platelet to lymphocyte ratio is a valuable prognostic biomarker in patients with colorectal cancer. Oncotarget. 2016; 7(18):25516-25527.
10. Song Y, Yang Y, Gao P, Chen X, Yu D, Xu Y, et al. The preoperative neutrophil to lymphocyte ratio is a superior indicator of prognosis compared with other inflammatory biomarkers in resectable colorectal cancer. BMC Cancer. 2017; 17(1):744.
11. Sipos F, Firneisz G, Muzes G. Therapeutic aspects of c-MYC signaling in inflammatory and cancerous colonic diseases. World J Gastroenterol. 2016; 22(35):7938-7950.
12. Terzic J, Grivennikov S, Karin E, Karin M. Inflammation and colon cancer. Gastroenterology. 2010; 138(6):2101-2114 e2105.
13. Ishii M, Wen H, Corsa CA, Liu T, Coelho AL, Allen RM, et al. Epigenetic regulation of the alternatively activated macrophage phenotype. Blood. 2009; 114(15):3244-3254.
14. Saeed S, Quintin J, Kerstens HH, Rao NA, Aghajanirefah A, Matarese F, et al. Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. Science. 2014; 345(6204):1251086.
15. Jain N, Shahal T, Gabrieli T, Gilat N, Torchinsky D, Michaeli Y, et al. Global modulation in DNA epigenetics during pro-inflammatory macrophage activation. Epigenetics. 2019:1-11.
16. Hanash S. Integrated global profiling of cancer. Nat Rev Cancer. 2004; 4(8):638-644.
17. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45(10):1113-1120.
18. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455(7216):1061-1068.
19. Kim D, Joung JG, Sohn KA, Shin H, Park YR, Ritchie MD, et al. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. J Am Med Inform Assoc. 2015; 22(1):109-120.
20. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HK, Frigessi A, Borresen-Dale AL. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014; 14(5):299-313.

21. Kim D, Shin H, Song YS, Kim JH. Synergistic effect of different levels of genomic data for cancer clinical outcome prediction. J Biomed Inform. 2012; 45(6):1191-1198.
22. Kim D, Li R, Lucas A, Verma SS, Dudek SM, Ritchie MD. Using knowledge-driven genomic interactions for multi-omics data analysis: metadimensional models for predicting clinical outcomes in ovarian carcinoma. J Am Med Inform Assoc. 2017; 24(3):577-587.
23. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015; 43(Database issue):D36-42.
24. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15(4):361-387.
25. T. H, B. L. On the exact distribution of maximally selected rank statistics. Journal Computational Statistics & Data Analysis 2003; 43(2):121-137.
26. Wong AK, Park CY, Greene CS, Bongo LA, Guan Y, Troyanskaya OG. IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Res. 2012; 40(Web Server issue):W484-490.
27. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015; 47(6):569-576.
28. Oliveira T, Silva A, Satoh K, Julian V, Leao P, Novais P. Survivability Prediction of Colorectal Cancer Patients: A System with Evolving Features for Continuous Improvement. Sensors (Basel). 2018; 18(9).
29. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 -[cited 2019 Aug 08]. Available from: https://www.ncbi.nlm.nih.gov/gene/.
30. Li J, Lan CN, Kong Y, Feng SS, Huang T. Identification and Analysis of Blood Gene Expression Signature for Osteoarthritis With Advanced Feature Selection Methods. Front Genet. 2018; 9:246.
31. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491(7422):119-124.
32. Gao L, Feng Y, Bowers R, Becker-Hapak M, Gardner J, Council L, et al. Ras-associated protein-1 regulates extracellular signal-regulated kinase activation and migration in melanoma cells: two processes important to melanoma tumorigenesis and metastasis. Cancer Res. 2006; 66(16):7880-7888.
33. Zhang L, Chenwei L, Mahmood R, van Golen K, Greenson J, Li G, et al. Identification of a putative tumor suppressor gene Rap1GAP in pancreatic cancer. Cancer Res. 2006; 66(2):898-906.
34. Li P, Wu YH, Zhu YT, Li MX, Pei HH. Requirement of Rab21 in LPS-induced TLR4 signaling and pro-inflammatory responses in macrophages and monocytes. Biochem Biophys Res Commun. 2019; 508(1):169-176.
35. Tang BL, Ng EL. Rabs and cancer cell motility. Cell Motil Cytoskeleton. 2009; 66(7):365-370.

36. Hooper S, Gaggioli C, Sahai E. A chemical biology screen reveals a role for Rab21-mediated control of actomyosin contractility in fibroblast-driven cancer invasion. Br J Cancer. 2010; 102(2):392-402.
37. Araki T, Milbrandt J. Ninjurin2, a novel homophilic adhesion molecule, is expressed in mature sensory and enteric neurons and promotes neurite outgrowth. J Neurosci. 2000; 20(1):187-195.
38. Ifergan I, Kebir H, Terouz S, Alvarez JI, Lecuyer MA, Gendron S, et al. Role of Ninjurin-1 in the migration of myeloid cells to central nervous system inflammatory lesions. Ann Neurol. 2011; 70(5):751-763.
39. Cho SJ, Rossi A, Jung YS, Yan W, Liu G, Zhang J, et al. Ninjurin1, a target of p53, regulates p53 expression and p53-dependent cell survival, senescence, and radiation-induced mortality. Proc Natl Acad Sci U S A. 2013; 110(23):9362-9367.
40. Langefeld CD, Ainsworth HC, Cunninghame Graham DS, Kelly JA, Comeau ME, Marion MC, et al. Transancestral mapping and genetic load in systemic lupus erythematosus. Nat Commun. 2017; 8:16021.
41. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet. 2010; 42(6):508-514.
42. Komatsu M, Li HO, Tsutsui H, Itakura K, Matsumura M, Yokoyama KK. MAZ, a Myc-associated zinc finger protein, is essential for the ME1a1-mediated expression of the c-myc gene during neuroectodermal differentiation of P19 cells. Oncogene. 1997; 15(10):1123-1131.
43. Smith DR, Goh HS. Overexpression of the c-myc proto-oncogene in colorectal carcinoma is associated with a reduced mortality that is abrogated by point mutation of the p53 tumor suppressor gene. Clin Cancer Res. 1996; 2(6):1049-1053.
44. Haberman Y, Karns R, Dexheimer PJ, Schirmer M, Somekh J, Jurickova I, et al. Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. Nat Commun. 2019; 10(1):38.
45. Ahechu P, Zozaya G, Marti P, Hernandez-Lizoain JL, Baixauli J, Unamuno X, et al. NLRP3 Inflammasome: A Possible Link Between Obesity-Associated Low-Grade Chronic Inflammation and Colorectal Cancer Development. Front Immunol. 2018; 9:2918.
46. Seo T, Tatsuguchi A, Shinji S, Yonezawa M, Mitsui K, Tanaka S, et al. Microsomal prostaglandin E synthase protein levels correlate with prognosis in colorectal cancer patients. Virchows Arch. 2009; 454(6):667-676.
47. Yoshimatsu K, Golijanin D, Paty PB, Soslow RA, Jakobsson PJ, DeLellis RA, et al. Inducible microsomal prostaglandin E synthase is overexpressed in colorectal adenomas and cancer. Clin Cancer Res. 2001; 7(12):3971-3976.
48. Williams AD, Korolkova OY, Sakwe AM, Geiger TM, James SD, Muldoon RL, et al. Human alpha defensin 5 is a candidate biomarker to delineate inflammatory bowel disease. PLoS One. 2017; 12(8):e0179710.

49. Alquraini A, Jamal M, Zhang L, Schmidt T, Jay GD, Elsaid KA. The autocrine role of proteoglycan-4 (PRG4) in modulating osteoarthritic synoviocyte proliferation and expression of matrix degrading enzymes. Arthritis Res Ther. 2017; 19(1):89.
50. Iqbal SM, Leonard C, Regmi SC, De Rantere D, Tailor P, Ren G, et al. Lubricin/Proteoglycan 4 binds to and regulates the activity of Toll-Like Receptors In Vitro. Sci Rep. 2016; 6:18910.
51. Sarkar A, Chanda A, Regmi SC, Karve K, Deng L, Jay GD, et al. Recombinant human PRG4 (rhPRG4) suppresses breast cancer cell invasion by inhibiting TGFbeta-Hyaluronan-CD44 signalling pathway. PLoS One. 2019; 14(7):e0219697.
52. Farwell SL, Kanyi D, Hamel M, Slee JB, Miller EA, Cipolle MD, et al. Heparin Decreases in Tumor Necrosis Factor alpha (TNFalpha)-induced Endothelial Stress Responses Require Transmembrane Protein 184A and Induction of Dual Specificity Phosphatase 1. J Biol Chem. 2016; 291(10):5342-5354.

Tables

Table 1. Demographic features of the study population

	Alive	Dead	P value
Age	62.9 ± 12.5	71.7 ± 12.6	<0.001
<65	111 (53.1%)	7 (18.4%)	
>=65	98 (46.9%)	31 (81.6%)	
Continuous	62.9 ± 12.5	71.7 ± 12.6	
Gender			0.407
Male	108 (51.7%)	23 (60.5%)	
Female	101 (48.3%)	15 (39.5%)	
Overall stage			0.201
Stage 1, 2	136 (65.1%)	20 (52.6%)	
Stage 3	73 (34.9%)	18 (47.4%)	
Stage 1	46 (22.0%)	2 (5.3%)	
Stage 2	90 (43.1%)	18 (47.4%)	
Stage 3	73 (34.9%)	18 (47.4%)	
T stage			0.046
T1, T2	50 (23.9%)	3 (7.9%)	
T3, T4	159 (76.1%)	35 (92.1%)	
T1	8 (3.8%)	1 (2.6%)	
T2	42 (20.1%)	2 (5.3%)	
T3	146 (69.9%)	32 (84.2%)	
T4	13 (6.2%)	3 (7.9%)	
N stage			0.113
LN negative	136 (65.1%)	19 (50.0%)	
LN positive	73 (34.9%)	19 (50.0%)	
N0	136 (65.1%)	19 (50.0%)	

N1	49 (23.4%)	9 (23.7%)	
N2	24 (11.5%)	10 (26.3%)	
Tumor location			1
Right colon	102 (50.2%)	17 (48.6%)	
Left colon	101 (49.8%)	18 (51.4%)	
Venous invasion			0.916
Negative	147 (80.8%)	25 (78.1%)	
Positive	35 (19.2%)	7 (21.9%)	
Lymphatic invasion			0.887
Negative	140 (75.7%)	24 (72.7%)	
Positive	45 (24.3%)	9 (27.3%)	
Follow up duration (months)	38.54 ± 31.26	39.97 ± 22.79	0.739

Table 2. Comparing the prediction performances in each model by concordance index

Features used	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Clinical features	0.318	0.648	0.726	0.706	0.791	0.921
Expression features	0.449	0.614	0.691	0.688	0.774	0.899
Methylation features	0.447	0.580	0.686	0.683	0.772	0.899
Expression features + Methylation features	0.337	0.647	0.727	0.715	0.826	0.884
Clinical features + Expression features	0.438	0.609	0.667	0.673	0.761	0.832
Clinical features + Methylation features	0.333	0.628	0.704	0.682	0.757	0.866
Clinical features + Expression features + Methylation features	0.326	0.655	**0.756**	0.708	0.818	0.883

Table 3. Association analysis between integrative model with clinical, gene expression and methylation features and overall survival

Hazard ratio (HR)	95% CI, lower	95% CI, upper	Z value	Adjusted P value	
Age	1.697	0.609	4.733	1.011	0.312
N stage	2.942	1.253	6.912	2.477	0.013
T stage	0.919	0.237	3.557	-0.122	0.903
Gender	0.910	0.390	2.121	-0.219	0.827
CEP250 (methylation)	0.592	0.364	0.963	-2.110	0.035
DEFA5 (expression)	0.786	0.462	1.337	-0.888	0.374
MAZ (methylation)	0.967	0.809	1.156	-0.369	0.712
NINJ1 (methylation)	1.339	0.910	1.968	1.482	0.138
NLRP14 (expression)	0.797	0.525	1.211	-1.063	0.288
PPARGC1A (expression)	0.808	0.636	1.027	-1.744	0.081
PRG4 (expression)	1.256	0.871	1.811	1.219	0.223
PTGES (expression)	1.399	0.937	2.087	1.644	0.100
RAB21 (methylation)	1.556	1.172	2.065	3.060	0.002
TERF2IP (expression)	1.452	0.893	2.360	1.502	0.133
TMEM184A (expression)	1.251	0.819	1.911	1.038	0.299
TNFRSF18 (methylation)	1.489	0.783	2.829	1.214	0.225
TNFSF12 (methylation)	1.132	0.801	1.600	0.704	0.481
TNPO3 (methylation)	1.465	1.092	1.967	2.543	0.011

Analysis was done with multivariate cox proportional hazard regression

CEP250, centrosomal Protein 250; DEFA5, defensin alpha 5; MAZ, MYC associated zinc finger protein; NINJ, Nerve injury-induced protein 1; NLRP14(expression), NLR family pyrin domain containing 14; PPARGC1A, Peroxisome proliferator-activated receptor γ coactivators 1 alpha; PRG4, proteoglycan 4; PTGES, prostaglandin E synthase; RAB21, Member RAS Oncogene Family; TERF2IP, TERF2 interacting

protein; TMEM184A, transmembrane protein 184A; TNFRSF18, TNF receptor superfamily member 18; TNFSF12 TNF superfamily member 12; TNPO3, transportin 3

Figures

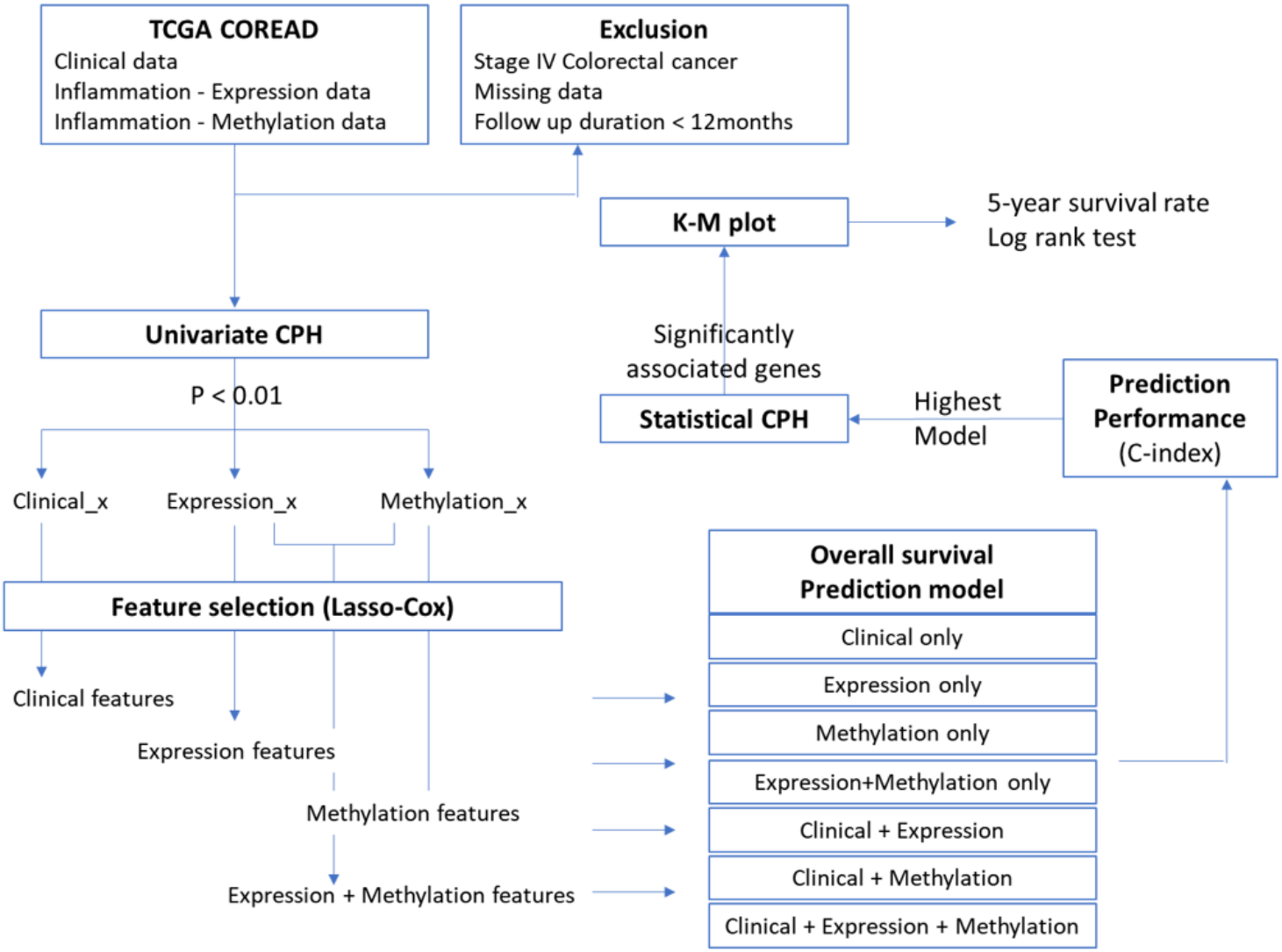


Figure 1

Overview of the analysis framework. With the gene sets related to inflammatory status and clinical information, we preprocessed them based on the significance of association with overall survival by performing univariate cox proportional hazard regression (CPH) analysis. Those features that passed the suggestive significance level (P < 0.01). Features with concatenated Expression and Methylation were also included. Feature selection was performed by using Lasso-penalized Cox proportional hazards regression (Lasso-Cox). Seven modes were designed by respective combination of features and the performances were evaluated by Lasso-Cox by C-index. Among the models, the highest C-index model went through multivariate CPH analysis for overall survival. Then, Kaplan-Meier analysis was done by the

genes which were significantly associated with overall survival and the log rank test was done for comparison of each group.

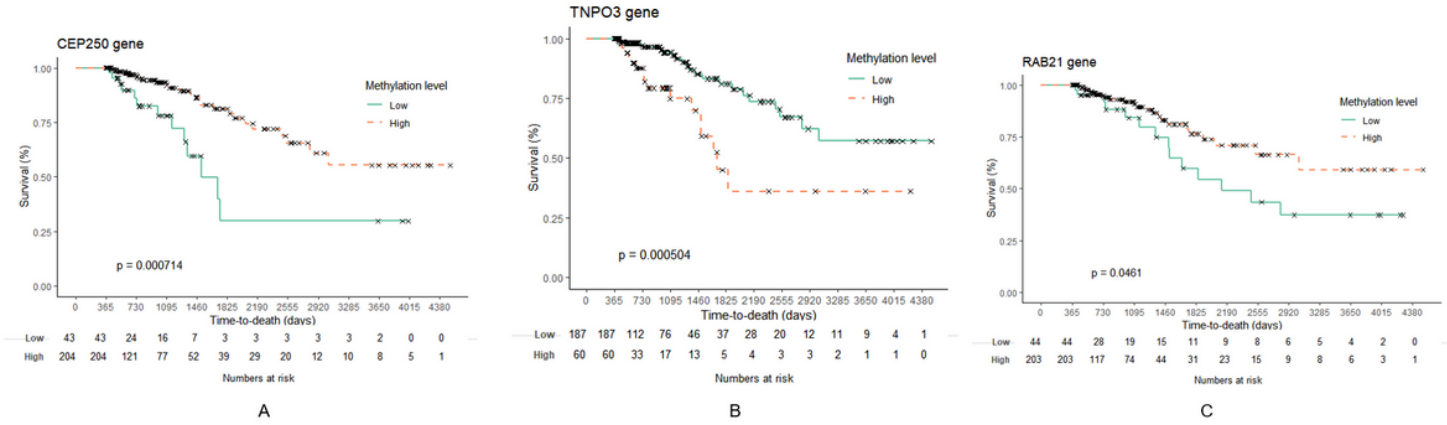
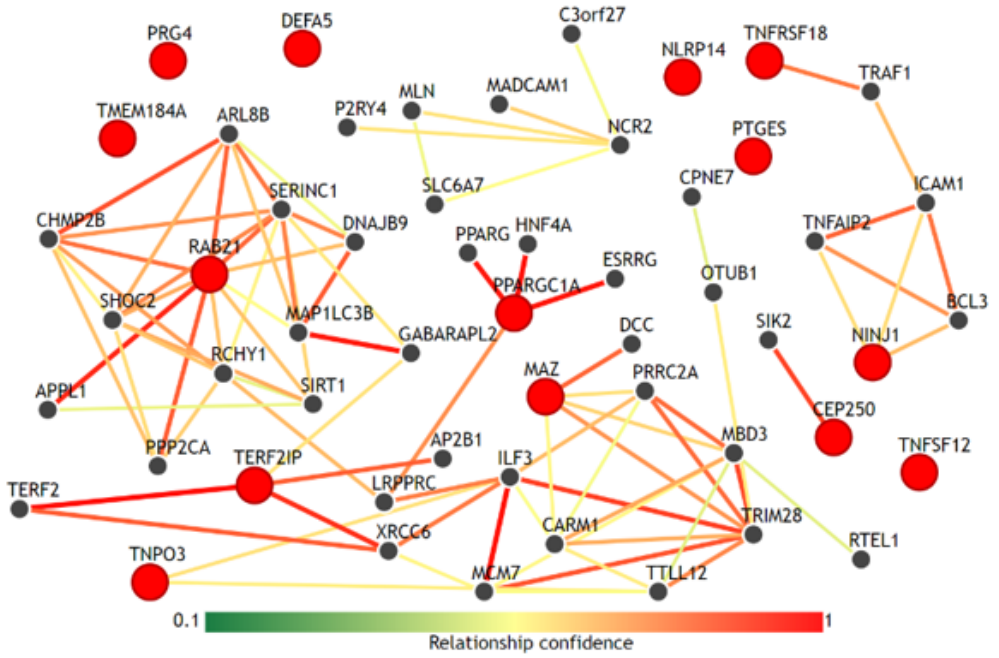


Figure 2

Kaplan-Meier plot for overall survival with genes significant in multivariate cox proportional hazard regression. Kaplan-Meier plot for overall survival with gene CEP250 (a), TNPO3 (b) and RAB21 (c). The patients were divided by the optimal cut offs for each feature based on MaxStat. (-0.923 for CEP250; 0.5 for TNPO3, and -0.367 for RAB21).



a

b

Figure 3

Functional network of the gene features used in integrative model with clinical, gene expression and methylation features for prediction of overall survival. a. Gene-gene networks queried by 14 genes in the model. Red dots are the queried genes (14 genes, Table S1) and the black dots are the gene predicted based functional networking analysis (0.5 confidence with 40 genes). Except for RAB21, TERF21P and PPARGC1A, most of the genes among 14 genes had less than 3 strong relationship with other genes by

network. b. Biological process networks queried by 3 genes (CEP250, TNPO3, and RAB21) that showed significant association with overall survival (confidence 0.5). All genes had unique biological process and there were no shared biological processes among three genes.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.docx